Using Excel to Support Data Analytics

Excellent, so welcome back to the last part of our introduction to data science.
So, stay our reduces talk a little about using Excel.
Excel is an excellent tool to kinda get your hands on a piece of data and data set.
And it's also a very commonly used tool throughout the business world.
So, I think there's gonna be this interesting transition from using Excel to Power BI which is gonna be the discussion of our next topic.
But Excel and Power BI are kind of synergizing where you're getting a lot of tools that are being developed for web applications, and pulling information from the cloud actually pulling into Excel.
Great, so
what we're gonna do in this demo is we're basically just build ourselves a histogram, talk a little bit more about what a histogram is.
We're gonna create a scatter plot, and
then we're gonna make a little of T-test.
So, hopefully it will be relatively painless, and should be fun.
Right, so the data set that we're gonna get is be gonna be a data set that I just randomly grabbed from the UCI machine learning repository.
And I know I mentioned cars multiple times in the earlier videos,
I'm actually not a car person, so
I really don't know too much about cars.
But that's what the data says it's gonna be on.
And just cuz I wanted to kinda talk about this data being kinda agnostic.
So, data science being data type agnostic.
And so, tool that you wanna apply at this level one stuff.
You just kinda apply to whatever data set you're looking at, and see what comes out, see how you wanna analyze it.
So, that's gonna be our approach.
So, let's go ahead and get started.
So, I'm just gonna go ahead and start up Excel.
And as it starts up, I'm just gonna create ourselves a blank workspace.
All right, great, so there, a blank workspace, I'll full screen it.
And there's a bunch of different things you can do.
What I'm gonna do is I'm just gonna go up to this Data piece up here.
Click on that, and I'm gonna go over.
And there's a bunch of different ways you can import data.
So obviously, we could just type in our data if we wanted to,
just like you normally would.
Maybe just make this a little bit bigger, so we can all see it.
Right, so we pull up our Excel spreadsheet here, and obviously,
we could just entered information like we normally would.
So we'll just be typing in something like ID, and then we'd say, 1.

And name, platypus, I don't know.
So, we may be talking about different pieces, but
a lot of data is already out there.
And so, as a data scientist it's not really our job to set up the data
structure, often we'll have the data structure already in place, and
we wanna be the ones who are manipulating it,
organizing it afterwards.
So, to do that, there's a bunch of different ways to import data.
And so, if you click on the Data tab, you can say, From Text, or
From HTML, or From FileMaker, or From a Database Query.
We're gonna go from From Text.
And I already downloaded the data set, and stuck it into a CSV file.
So, here it is, it's called car_data.csv.
Get data.
Now, we talked about CSV files, and
CSV files are just these comma-separated value files.
And so, if we look at it,
it's really kind of an ugly looking files.
So it's basically columns that are designated by commas.
And so, it might have this thing that says delimited, and
that's basically how they're spaced apart, and so
if you just hit Next after this, great.
And it says, how are they split apart?
And so, you'll see in the field as you're going through this,
what the data looks like after its been parsed.
So, if we leave it as just tab, it'll look still pretty ugly, but
see how it's a comma-separated value.
So, perhaps our delimiter is a comma.
So, if we click on comma, we'll notice that it pops out, and
starts to look more like a table.
And so, that's pretty much what we want our data to look like.
So let's just go ahead, and hit Finish, and
it'll asks us where we wanna put it.
So, let's put it into starting from A, and
starting from actually 2 cuz we're gonna give this a bunch of titles.
Cool, and so our data just got pulled in.
Fantastic, so now we're sitting, and
we have a bunch of different data from 1985 car automobiles.
And so, what I would recommend if you want to kind of
work along with this demo is go ahead, and get that data set
from UCI, pull it in, and just follow on.
So, you can always pause the video, you can always stop it, and
kind of make sure that you're at where we are, and then work with it.
Because kind of seeing somebody do it is useful, but
doing it yourself is much, much more useful.
So, all right, so now we've got this, but

we also have some information about the text file that we're using.
So, everything from where it came from originally, how
to cite it, and information about what the different columns are.
And so, we're gonna use that information about what the different
columns are, and just fill that in, so
that we have a table that kind of is much more humanly readable.
And so, again,
we're just generating this level one data that we've talked about.
So, we want this kind of humanly readable sort of data set that we
can kind of deal with, and actually makes sense of.
So, we're just gonna grab that set.
Actually, I'm just gonna type it all in.
So it's kinda take me a few minutes, so feel free to speed up the,
The demo, or if it's going too fast for you still, pause the demo, and
just enter it in yourself if you're working through.
So, let me just show you how I'm gonna do that.
So, I'm just gonna basically split my screen, and
so that I have both pieces of information.
Now you just kinda see what I'm doing, while I'm doing it.
All right, so first one, symboling, and
we're just gonna go all the way through price.
Great, so hopefully, everything has all that entered, so
that they're relatively happy with the data.
And what we're gonna do next is we're just gonna organize that
data sets, a little bit more, I don't know, it's just better.
[LAUGH] We're just gonna create it into a table.
There we're just gonna select all the data.
And basically, you just say, hey, why don't we just,
I insert this thing.
I'll just close this down, sorry.
We've got everything, all the information from our text file, so
let's close that.
And if you just click on Insert and hit Table,
it'll reformat that, and ask you where you want your table from.
All right, so A1 perfect, Z1 or Z2 without 206, great.
And my table does have headers, now you just hit, OK.
Yes, and much better.
So now we have, so it's a little bit more readable.
And one of the cool things about converting something to a table
using Excel, it's now sortable.
And so, for example, if we wanted to sort something by the number of
doors, we can go, and get all of the two components.
So we'd say, something like two and Enter.
And so notice, the number of doors now is filtered,
going by two all the way down.
Then it goes to four after that.

Great, so, the tables just give you a little bit more
access to being able to do things with your data.
All right, so as we mentioned before,
one of the really critical pieces of dealing with data in a,
kind of getting and understanding of data is dealing with the histogram.
And kind of looking at just histograms in general,
to kind of get us a feel for how our data are arranged.
Right, so now, we have all of our data with nice columns, and
we have a nice table.
So, at this point, let's go ahead and build our histogram.
So, why don't we take this bore, and use that in C, what it looks like.
So, as I mentioned, I don't really know too much about vehicles.
So, we're just gonna choose one, and see what that data look like.
So, first thing we wanna do, because it's kinda this level one data where
it's kinda humanly readable, we're just gonna scroll through this data.
And look at kinda the range of data that are in bore, so
that we can generate these different bins.
So we can kind of parse that data,
and organize it in a way that makes sense.
So, let's go ahead and take a look at it.
It looks like we have a range that's gonna be between, 3 or
2.5 maybe, and Looks like about four,
and so we're just gonna go by
intervals of a tenth and create some bins.
So all you gotta do is go over and create another column.
I'm just gonna label this column bins, and we'll start off at 3.5.
Let me just make sure that's a good range.
3.4, 3.1 so actually that looks like we can go all the way down to 2.5.
So let's just go ahead and start at 2.5 and just work our way down.
2.5 and obviously we could just do this.
Just write a quick little script to do that.
But I'm just gonna do it manually cuz I'm gonna do it that way.
Great, so now we enter our bins and we have a range of bins.
But now, what we're gonna do is basically ask for
Excel to basically organize this data into bins in between.
So one bin will be 2.5 to 2.6.
The next bin will be 2.6 to 2.7.
And it will just count how many elements in our
four column fit into those criteria.
And it'll generate ourselves a nice little graph, all right?
So how do we do that?
Well, we go over to data solver and we say, go to histogram, right?
So click on histogram, perfect.
And we ask what columns we wanna do it with.
All right, so we've got AB and we want our bins from all of AB.
So bin range, great, look at AB.

And then our bore is gonna be column S, so we just click on that.
Great, and let's just hit OK, see what happens.
No, histogram input range contains non-numeric data, right.
So something's going wrong.
So somewhere in our data set we have data that are not accurate,
or not viable to us.
So let's cancel that and go ahead and look through our data.
So we look at our bore column, cruising along, cruising along.
That looks good, looks good.
No, there's some weird question marks here.
So some of the automobile manufacturers when this data set
were made didn't published their bore.
So, what do we want to do with these data?
Well, we need to either get rid of them in some way or
give them some of value that we can kind of deal with.
So, our bins range from 2.5 to 4.
And so if we put these data outside of the bin range, but
make them numerical.
So Excel at that point will know what to do with them.
It'll look at them and be like do you go into a bin? No.
But at least I can read you,
rather than being trying to read a question mark.
All right, so let's make sure that there aren't any more in there.
And we're just gonna rerun our histogram.
Great, so data analysis, histogram, and just to make sure there are no
possible chances that we're catching the title we're gonna start from s2.
That's gonna be starting from there and
we're gonna all the way down to s206.
And same with ABs we are gonna go from AB2 and
then we're gonna go all the way down to AB19.
All right.
So, let's go ahead and make ourselves a chart style output.
And there's a couple of other types of histograms that one can look at
array those and similar percentages.
Feel free to explore this on your own.
So basically these are some sum-mated histograms and
ways to look at data as they kind of rise over time or
over whatever type of input you have on your x value.
So your bins.
And so basically you're gonna end up with a percentage as they increase.
So, but we're not gonna worry about those for now.
So let's go ahead and hit new chart and see what we get.
Alright, cool.
So there's our nice little histogram.
And it doesn't look particularly abnormal, but it doesn't
look like a normal curve either, so there's kind of these definite

outliers down at this low range, some outliers at this high range.
And there is sort of a sloping peak,
but I wouldn't consider this necessarily a normal distribution.
There are enough data points in here, so
200 data points, that you should be able to see some sort of
tendency towards the mean if you have a normal distribution.
And this is one of the reasons that when we mention kind of low or
kind of level one data, that analysis can be challenging.
If you only have a few data points, maybe you have 30 data points you
may not have enough data to actually determine whether or
not you have any normalcy between or
generate a normal curve in your data when you look at it.
But with 200 data points, you should start to see something that starts
to look normal, if it is actually normal.
And this, just my data intuition, says this is probably not normal.
And we kind of talked about some of those reasons why before, so
you basically have these different pressures in the terms of
manufacturing.
Vehicles are limited by how big they can be and how much compartment
space there are for the engine and so you are kind of limited as to
how you can generate and what bore size you can generate.
In addition, you're gonna be having conversations with other
manufacturers as who's generating piping for your system.
As a car manufacturer, you're probably getting that from somewhere
else, so they have their own constraints.
And so the majority of car manufacturers are gonna get their
materials from a few locations and so all these different constraints,
but it's good to understand what our data look like.
All right, awesome.
So that's part one,
just kind of getting a sense of our data using a histogram.
I highly encourage you to do this for
pretty much any type of data set that that you're looking at,
you basically wanna get some understanding of it.
So let's go ahead and do a little bit of data analysis now.
So one of the types of data analysis that we might wanna do is
understand whether or not there's a correlation between
two different events in our data set.
So we can go back over to our data and take a look at a couple things.
And we're gonna choose something that I
think should have a correlation.
So of our data that we have here,
well we've got things like horse power and engine size.
Okay, so that makes sense that there's probably a correlation
between those two things.
So what is a correlation mean?

So a correlation means that as you go up in one thing,
you go up in the other thing.
Or if you go down in one thing, you go down in the other thing.
And so, one event is correlated with the other one, right?
And so, I think many of you have heard this kind of term,
correlation does not necessarily imply causation.
So, there's definitely, it's a trend, not a causal relationship.
So a goes up, b goes up sort of event.
Why, who knows at this point.
So let's go ahead and take a look at that,
and so how we would do this is we will generate a scatter plot.
And so, again, a lot of these tools are really common and so Excel
pretty much has them all set up for us, and so we can just go over to
insert and then hey look, there's all bunch of recommended charts.
Wow.
All right, so scatter plot, xy scatter, great.
Let's just go ahead and click on that.
And there's a scatter plot, fantastic.
And we get this nice little chart that's already pre-made for us.
We wanna change our data so
we're gonna click on this thing that says select data up here.
So just hit Select Data, and we can just choose what we want to put in.
So, things that we want to put in,
we said we wanted to go with engine size, and horsepower.
So figure out where engine size is.
Where did I put engine size?
Engine size, here we go, Q.
Great, and then horsepower, great.
And so, Excel will just basically, you click on the column type and
if it's in this table format it knows that it will
get rid of the header for us.
Let's just subtract a bunch of these things.
There we go, perfect.
And hit OK.
See what it looks like. Great.
So, now we have this chart of
different- whoops, here.
I meant to grab the thing and make it bigger.
Re-center that. All right, so
now we have all of our data points sitting here on this nice chart.
So on the y axis, we have engine size, and
then in the x axis we have horsepower, all right.
So let's take a look at this and think about what it means.
So it looks like there's this general trend that's kind of
going upwards, right?
So our dots have some sort of potential trend.

So how we actually look at that, so basically
what we wanna do if we're trying to look at correlation between data and
between two different axis, is we just wanna say hey
what's the best fit line for this is a really common way to do that.
And so, again, this is kind of a pre-built strategy, so
if you just right click or
control click onto one of the dots, doesn't matter which one.
It will ask you, hey what do you wanna with this thing?
Well you can label it or whatever.
But all we wanna do is add a trend line.
All right, great so XLR already knows how to do this.
You can dictate what type of trend line that you want to use.
Linear is a good starting point.
But you could play around and do a polynomial fit and
see if it fits best.
You know, whatever you wanna do.
But as you understand what those different trend lines mean,
or why you might want to use them, for example, you might want to use
a logarithmic inline if you think that there's some sort of event
that's gonna cause some sort of asymptotic relationship.
So basically, as you go up in one structure, the, if you go up in A,
eventually B is no longer going to increase but
it's still gonna correlate along this line, so, no matter how big you
get the engine, your horse power is not gonna go up or vice versa.
All right, so we've got this trend line, and then what we wanna
do is we wanna ask it to display a couple of things, so
displaying the equation, as well as the R squared value.
So the R squared value is basically how well does this thing correlate,
so how much do these plots, or these dots, scatter from this line?
If something, every time you went up in one, in type A, and it went up,
always went up exactly one in type B, you have a perfect correlation,
so every single line would lie on that dot, right?
Or every dot would lie on that line, sorry, but most things
don't have a distribution that's perfectly correlated.
So, here we have a correlation, and it tells me, as I said,
a value of one means that it's perfectly correlated,
a value of zero means it's uncorrelated.
All right, we also have a slope, and so,
the slope of our line is positive, so where lines have slopes, and
lines have this equation, mx plus b, so where m is the slope of the y and
b is the y intercept.
So, the slope of that line is basically gonna tell us whether or
not the correlation is positive or negative, so you could have slope
that was every time you went up one in A, you went down one in B, right?
So that'd be a slope of a negative slope.
All right, so, awesome, so this is another piece of data analysis.

So if you were a car manufacturer, and you're like, okay, so,
we really wanna build a car that has 300 horsepower, and
we're trying to figure out how much material we're gonna need, and
kind of what type of engine we're gonna need, one might build this
type of data, and look at what cars are already out there, and say,
all right, so we're gonna probably be estimating something in the range
of 200 to 300, I guess, we look at the line here.
So here's 300 horsepower, and somewhere in around the 245 ish
range, so we can plug it into our equation and
actually figure out the exact value if we wanted to.
All right, so that's the value of trend lines, again,
a super common application, and one of the cool things about trend
lines is it's a really common first machine learning strategy.
So basically,
one way to machine learn stuff is basically do all these calculations
so you calculate a line and be like, okay, so how far is this dot from
that line, how far is this dot from from this line,
how far is that dot from the line, et cetera?
And so, you can throw that into a computer program and
you can just start drawing lines, and
so, you draw your first line straight across,
and then be like, these dots are really far away from it on average.
And then, you start to just change your line around, just doing this
kind of hypothesis iteration sort of strategy until you have a best fit.
So, something to kind of think about, using this understanding of
what these trend lines work, or how these work,
to kind of get yourself kind of this core basic into some of the machine
learning structures that you're gonna maybe wanna dig into later
as you become more of an experienced data scientist.
All right, cool, so there's two different techniques that we've look
at, the last thing I wanna do is I wanna run the t-test, and so,
what is a t-test?
A lot of people kind of get lost on t-tests, so
I'm not gonna go too far into it, but basically it's a way to
compare averages between two different sample sets.
And depending on how your data are distributed,
if they're normal, you might wanna use a student t-test, or
you might use a Welch's t-test, or these different structures,
depending on what your data look like to kind of figure out whether
or not two data sets are actually the same type of thing, or
if they're two separate types of things, is a really common use.
And so, maybe we wanna go ahead and take some data and
look at number of car doors, and, I don't know, wheel base length.
And our null hypothesis in this case is gonna be something to the effect
of, hey, it doesn't matter how many car doors you have,
because wheel base, or wheel length, is gonna be not correlated with it,

so, they could both be from the same data set, so let's go ahead and
take a look at that, and see how we would attack that using Excel.
I'm just gonna delete this guy, now we know our scanner plot, all right,
so how I'm gonna do this, we already sorted this by number of doors,
and so, let's just go ahead and grab our Wheel_Base and
go all the way down until we get to four doors.
All right, cool.
So, I'm just going to move this over into another column.
Great, so there's Wheel_Base, we're just gonna name this
part Wheel_Base, and we'll call it Wheel_Base 2 Door.
Great, and then I'm gonna grab the four door variant, so
just going back over.
All right, Wheel_Base, and it was 90 ish.
All right, just go down, grab all these guys, oops, come on, scroll.
Lots of four-wheeled, four-doored cars, and
there's a couple of cars that have questionable numbers of doors.
I'm gonna go with their Jeeps, that I actually don't know, [LAUGH],
a Dodge and a Mazda, why don't they know how many doors they have?
Doesn't make sense, but whatever, right?
All right, so, we'll just go ahead and move that over there,
we'll call this one Wheel_Base 4 Door.
All right, great, so now we have two data sets that we wanna compare
whether or not they're the same or different.
So our null hypothesis is that it doesn't matter how many door that
we have are that they are going to fit and look the same,
and so, let's go ahead and do this.
So now, we're just gonna go over to data, so again,
go to our data analysis, hunt down, so, we're gonna say, hey,
our assumption is that we're gonna make have t-test and
we're gonna assume unequal variances.
Great, all right, and just so you guys know, I'm gonna be designing
a little bit, this whole part is a little bit duplicitous, so
just be aware that something else is gonna happen at the end of this.
All right.
Cool, so, variable range one and variable range two, great.
All right, so we can hypothesize that there's gonna be some
difference between the two of them to start with, right now,
our hypothesis is that they're exactly the same, all right, and
then let's just run it, so we're gonna create it as a new worksheet.
Cool, all right, so, we get this data back, and what this data
looks like, it is to tell us the means of the two variables, so,
the 2 Door sample has a mean of 95, the 4 Door sample has a mean of 101,
and also how many samples are in each group?
So, there's 89 samples and a 149 samples in the two different groups,
and once you start to get bigger sample sizes, again, it's easier to
kind of look at the means, so the more samples you have, the more

tendency there is to kind of go to what they're supposed to be.
And so, if you had a million samples or an infinite number of samples,
the mean would be whatever the population is, right?
So if your sample looked at everything in the population,
you would know exactly whether or not that thing is the mean or
not, but right now, we're taking up sub-populations of a sample from
that population and saying, this could be the mean, this couldn't
be the mean, so, we kinda wanna know how many observations are.
So, we get these values down here, so these values are these p-tests,
so these p-tests, or these p values,
are basically the probability that these two things are the same,
and notice that they're ten to the negative 12th ranges.
And so, what that's basically telling us is that there's a really,
really small chance that if you just grabbed two random
groups of the same type of object that you'd end up with
these distributions, all right, now, why?
So that's really cool, it's a super useful tool, t-tests are great,
definitely take some time and play around with some different t-tests,
read up about them when you're done with this video.
But, what I did that was duplicitous here is we didn't look at
the histogram of these things, and the way t-tests are implemented on
Excel are they're based on these student t-tests, and
they need normal distributions, and so we don't know these things were
normal distributions before we actually attacked it.
And, in fact, we have some evidence that car manufacturing is
not normal, and so this is kind of the example of knowing the tools
that you're using, and jumping in and really addressing some of
these really interesting problems is really fun.
So, go ahead and jump into machine learning, but don't forget to really
emphasize the basics, so you don't make a pitfall and make a mistake,
like I just did, where you're saying, hey, this is what these
analyses tell us, when that's not really what these analyses tell us,
so, hopefully that was all helpful.
And, with that, we're really excited to have had you join us for
this program, and hopefully, you'll stick with us through Power BI and
then integrating R with Power BI.
But if not, there are some really excellent tools
to kind of expand your data science horizons, kind of in a structured
format where you're learning the statistical core.
Going through and learning how to program in Python,
learning how to program in R, and then applying machine learning
values from that point, cuz the goal of this is really to kind of give
you the overview of what data science is, so hopefully,
you enjoyed it, and I definitely enjoyed having you here.